**1.Data Manipulation**

In the beginning of this problem, we read in data *AngleClosure.csv*, delete the columns corresponding to factor variables EYE, GENDER, ETHNIC, HGT, WT, ASPH,

ACYL, SE, AXL, CACD, AGE, CCT.OD, and PCCURV mm, and then delete rows of the dataset which have any missing values

|  |  |  |
| --- | --- | --- |
|  | Observations | Predictors |
| Read The Data | 1468 | 24 |
| Omit specific attribute | 1468 | 11 |
| Delete missing value row | 1468 | 11 |

**2.Develop Prediction Models**

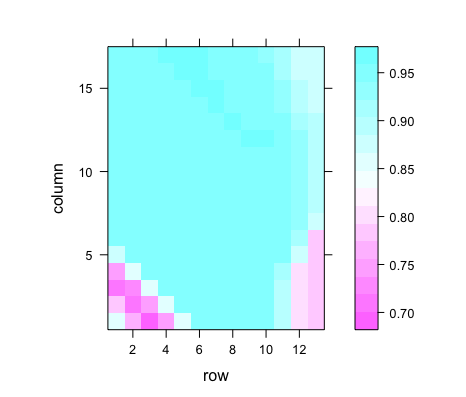
Five prediction models are chosen, and they are

* Support vector machine (e1071)
* Neural network (nnet)
* Random forest (randomForest)
* Boosted model (ada)
* Logistic regression

**3.Model Parameter Selection**

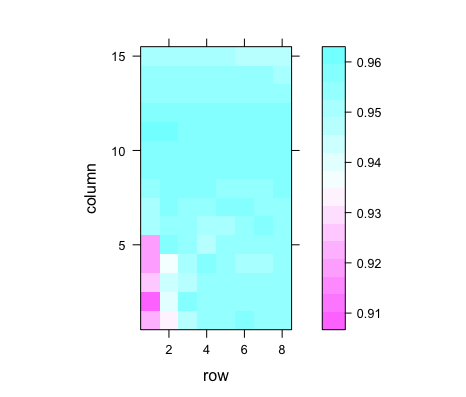
Support vector machine:

There are two parameters that need to be tuned in svm: *gamma* and *cost*. Therefore, we use a grid search to find the best combination of the parameters.

When Gamma=0.001, Cost=3.162, the max average auc is 0.95903

Neural network:

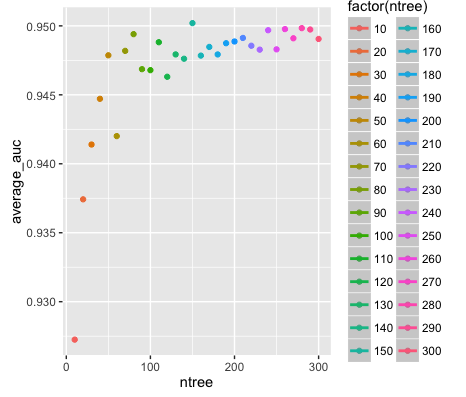
There are also two parameters that need to be tuned for neural network: the size and the decay. we use a grid search to find the best combination of the parameters.



When Size = 3, Decay = 0.1, the max average auc is 0.95957

Random forest:

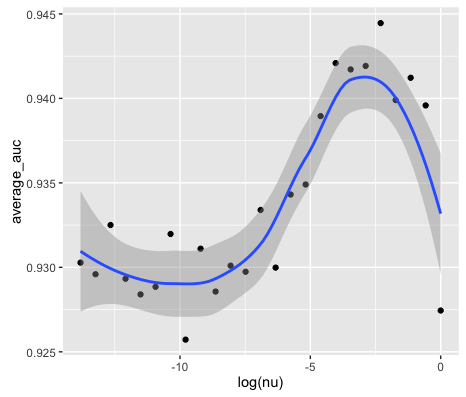
In random forest, the only parameter that could be tuned is the number of trees.



When number of tree = 150, the max average auc is 0.950198

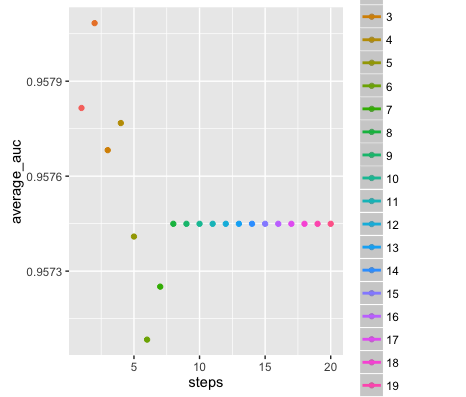
Boosted model:

In Adaptive Boosting method, the parameter that need to be tuned there is nu.



When nu = 0.1, the max average auc is 0.94445

Logistic regression:

In Logistic Regression, the tuning parameter is the steps numbers allowed in the Logistic Regression Model using two-directional stepwise AIC regression.

When steps = 2, the max average auc is 0.95808

**4.Stacking**

We use 10-fold cross validation with the selected models on the original training data. We predicted each model for the testing fold data and then store with corresponding response. Using store data for constrained and unconstrained optimizations.

The stacking models with and without constraints are shown in table below.

|  |  |  |
| --- | --- | --- |
| Model | Constraint w | Unconstraint w |
| Support vector machine | 0.1457315 | 0.15072539 |
| Neural network | 0.2662107 | 0.27214284 |
| Random forest | 4.312247e-19 | -0.06550852 |
| Boosted model | 0.1891759 | 0.21353243 |
| Logistic regression | 0.3988819 | 0.41952495 |

**5.Validation**

We manipulated the Cases and Controls data to ensure we only have the necessary columns and that we preferentially choose right eye data over left, and try to test the performance of the trained 7 models.

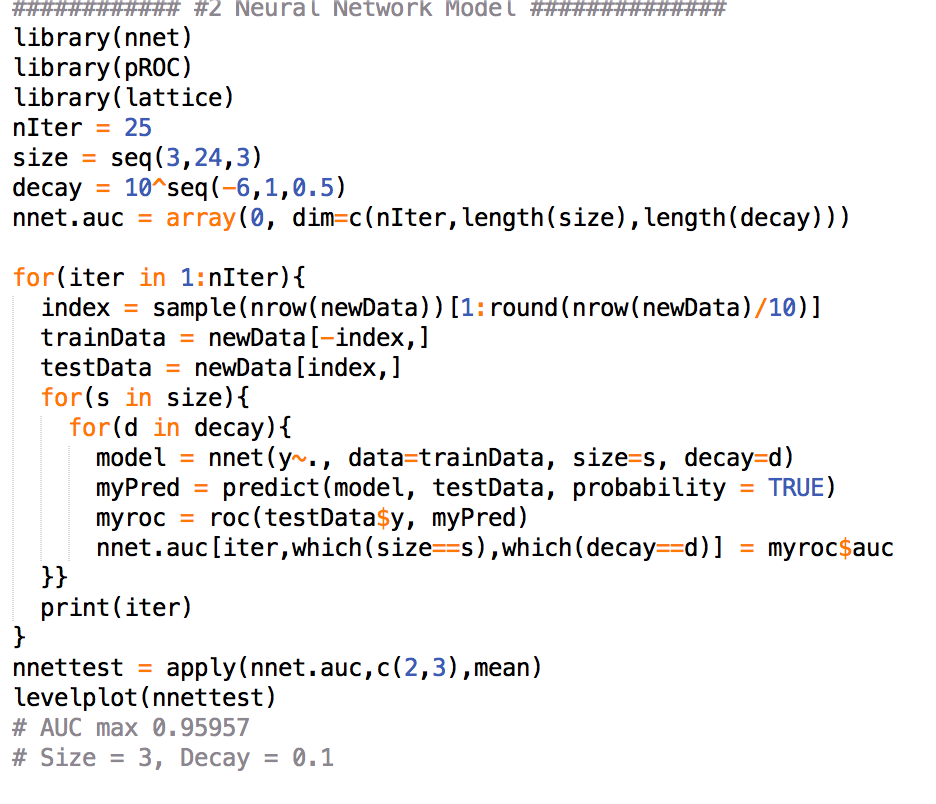
|  |  |
| --- | --- |
| Model | AUC |
| Support vector machine | 0.9519 |
| Neural network | **0.9649** |
| Random forest | 0.952 |
| Boosted model | 0.961 |
| Logistic regression | 0.9536 |
| Stacked Constrained | 0.9597 |
| Stacked Unconstrained | 0.959 |

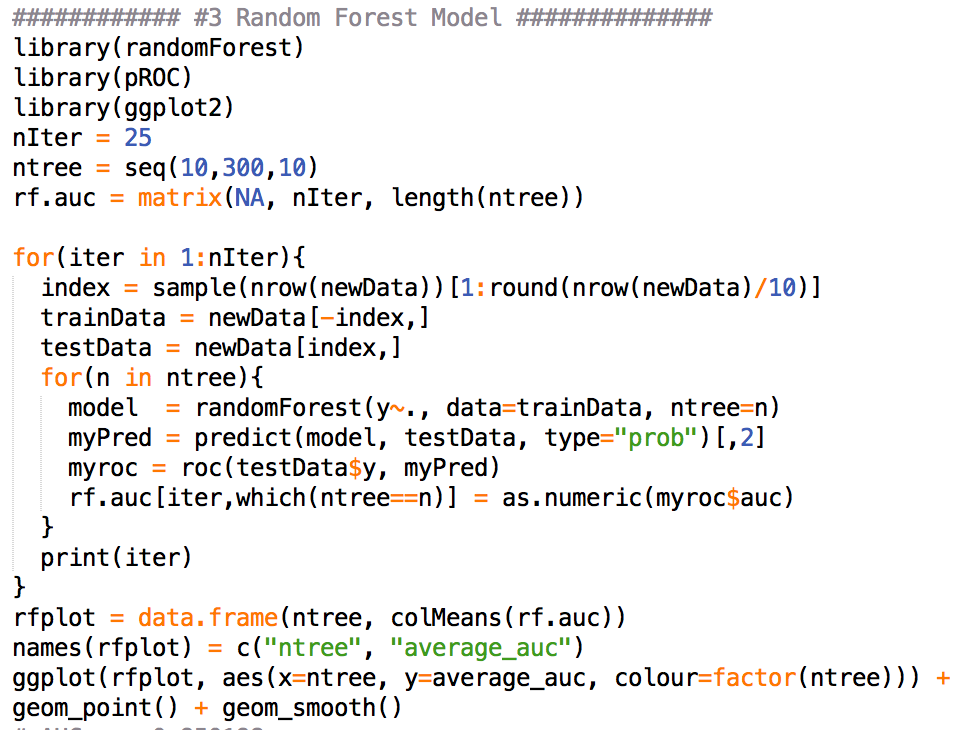
As we can see, the best model generated on the validation data set was the **Neural Network with an AUC of 0.9649.**

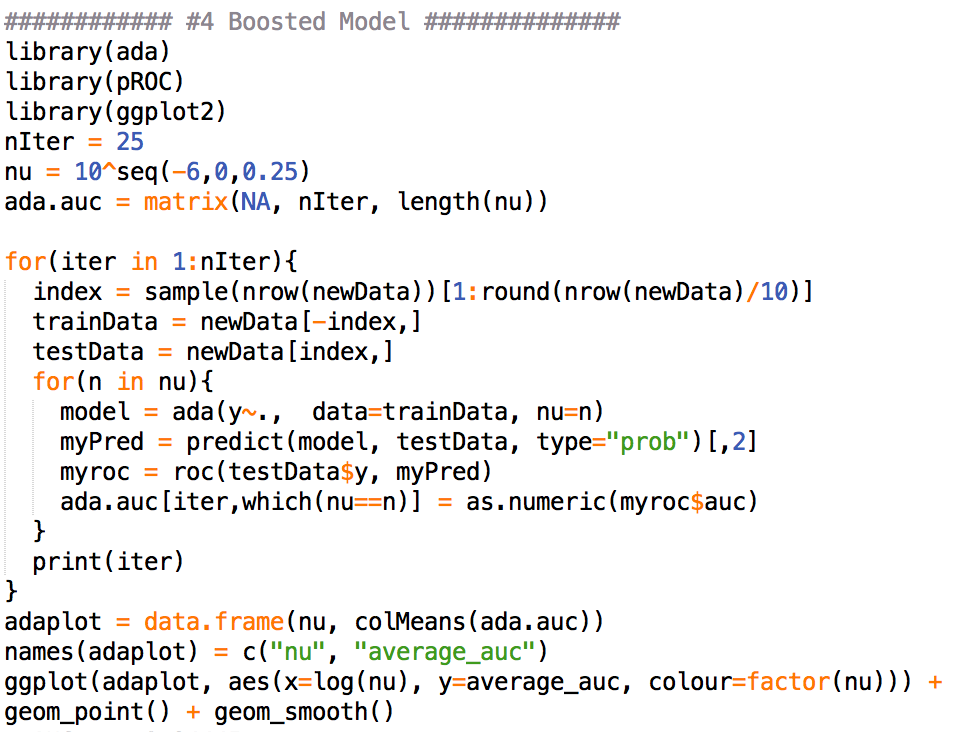
**6.Visulazation**

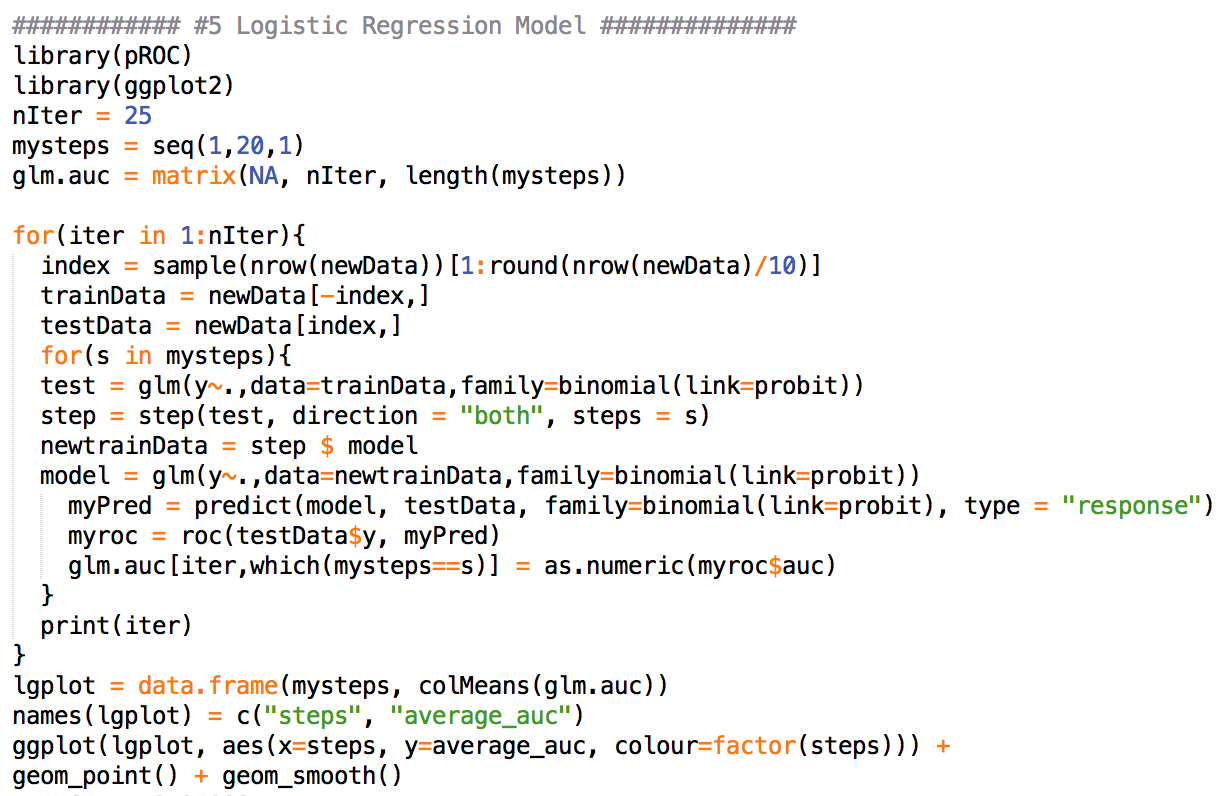
7 models’ ROC Curve.

|  |  |
| --- | --- |
| SVM: AUC = 0.9519 | Neural Network: AUC = 0. 9649 |
|  |  |
| Random Forest: AUC = 0. 952 | Ada Boost: AUC = 0. 961 |
|  |  |
| Logistic Regression: AUC = 0. 9536 |  |
|  |  |
| Stacked Constraint: AUC = 0. 9597 | Stacked Unonstraint: AUC = 0. 959 |
|  |  |

R CODE

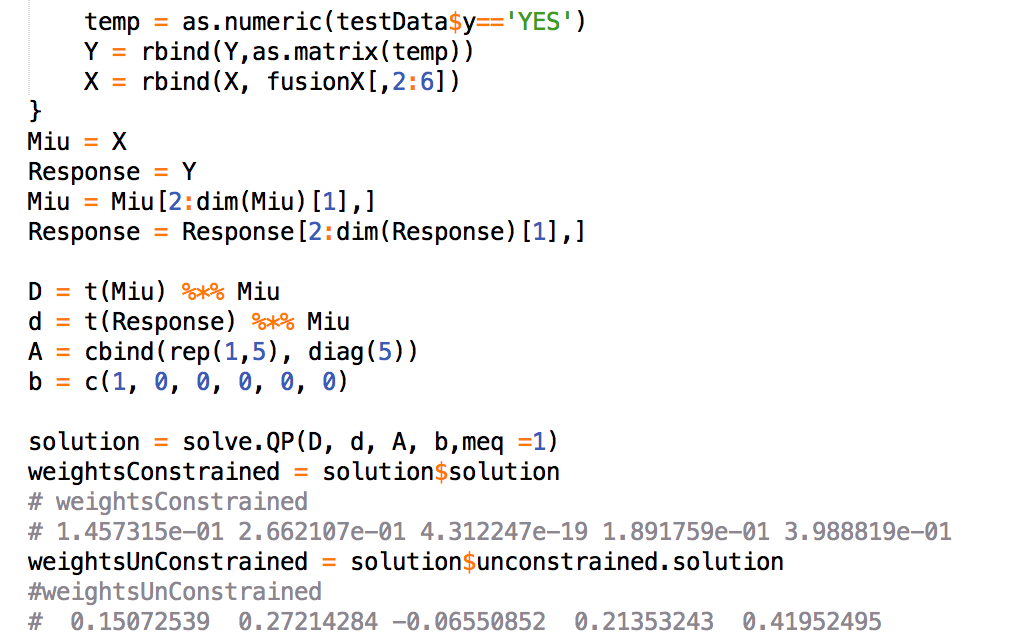






Stacking





Validation



